DOCUMENT RESUME

ED 399 267 TM 025 352

AUTHOR Ostrander, Laura R.
TITLE Multiple Judges of Teacher Effectiveness: Comparing Teacher Self-Assessments with the Perceptions of Principals, Students, and Parents.
PUB DATE 9 Apr 96
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Communication (Thought Transfer); Educational Environment; Elementary Secondary Education; *Evaluation Methods; Grading; Homework; Instructional Effectiveness; Parent Attitudes; *Principals; *Self Evaluation (Individuals); *Student Attitudes; *Teacher Evaluation; *Test Construction

ABSTRACT

While teachers traditionally have been evaluated by school administrators, the call for multiple and variable lines of evidence in teacher evaluation has had many proponents. To evaluate the use of multiple judges, teacher ratings from four judges using a common evaluation instrument were compared for 93 teachers from grades three and above. Thirteen to 15 students assigned to each teacher and 6 to 18 parents of their students complete evaluations. The teachers (n=108) themselves and the principals of their schools also participated. Subscores on the prepared evaluation instrument were determined for each teacher in the areas of classroom environment, grading, homework, communication, instruction, and interpersonal relationships. The highest ratings were given by principals, and the second highest ratings were by teachers themselves, although the population means provided by teachers only differed significantly from those of the principal in the homework subcategory. However, the correspondence between these two sets of ratings was low. Teacher ratings by parents and students, while quite high, were lower than those given by teachers and principals in each of the six areas, and teachers were viewed less positively by students than by parents. Findings suggest that use of multiple judges may provide unique perspectives on teacher performance, resulting in fairer and more comprehensive evaluations. (Contains 1 figure, 16 tables, and 37 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
LAURA R. OSTRANDER

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 399 267

MULTIPLE JUDGES OF TEACHER EFFECTIVENESS: COMPARING TEACHER SELF-ASSESSMENTS WITH THE PERCEPTIONS OF PRINCIPALS, STUDENTS, AND PARENTS

1996 AERA Annual Meeting

New York

Laura R. Ostrander
Virginia Beach City Public Schools
3512 George Mason Drive
Virginia Beach, Virginia 23456

April 9, 1996

TM025352

BEST COPY AVAILABLE

INTRODUCTION

In spite of increased knowledge about effective teaching and about successful organizational improvement, the model used for assessing the job performance of classroom teachers has changed little in this century. Current practices do not reflect developments in the field nor do they support recent reform efforts (McGreal, 1994; Ellett & Garland, 1987). Evaluation commonly consists of a tenured teacher being observed for 20 to 30 minutes once every two or three years. Sometimes a conference follows the classroom observation, which may or may not be announced in advanced. Nontenured teachers are evaluated in much the same manner but usually more often (Haefele, 1993). This model for evaluating teachers has serious shortcomings (McGreal, 1994; Medley, Coker, & Soar, 1984; Scriven, 1981), but it is entrenched in American schools.

The primary evaluator for teachers has traditionally been the principal. While evaluation is commonly based on observing the teacher in the classroom, it is influenced by interactions with the teacher during other activities (Scriven, 1981; Stodolsky, 1984). Few view this practice as successful in bringing about either individual or institutional improvement (Frase & Streshly, 1994; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). The evaluation of teacher performance is seen by many as the single most important responsibility of principals, but it is viewed as a process overdue for improvement (Larson, 1984; Haefele, 1981).

The traditional responsibility of principals as performance assessors and as decision makers about teacher quality has many critics. Charges of bias, prejudice, cronyism, favoritism, patronage, lack of expertise, and insufficient

training account for some of the problems (Epstein, 1985; Scriven, 1981; Stodolsky, 1984). Additionally, the work demands of the principalship do not permit building administrators to devote enough time to the process for teachers to have confidence in their ratings (Haefele, 1992). Principals tend to view the responsibility for performance assessment as burdensome (Frase & Streshly, 1994; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984), and teachers view the process as perfunctory (Manning, 1988).

There are two important arguments against continued reliance on the principal as the sole evaluator of teaching quality. First, the practice places the teacher's fate in the hands of a single judge (Peterson, 1987; Harris, 1986), and second, it is a model that no longer reflects the way American schools operate (McGreal, 1994; Ellet & Garland, 1987). Recent reform movements that foster stakeholder empowerment, collaboration, and increased teacher accountability require better ways to validate classroom effectiveness (Dixon, 1994).

Several have called for the development of multiple and variable lines of evidence about teacher performance to improve the state of teacher evaluation (Harris, 1986; Darling-Hammond, Wise, & Pease, 1983; Peterson, 1987; Stanley & Popham, 1988). This movement parallels the national trend toward increased client involvement in school governance and decision making. The knowledge base regarding effective multiple lines of evidence is, however, inadequate, and many potential data sources are relatively untested (Scriven, 1981; Peterson, Gunne, Miller, & Rivera, 1984; Epstein, 1985; Peterson, 1987).

Follman (1992) declared that there is "no one uniquely valid" data source on which to base teacher evaluation (p. 169). Peterson, Gunne, Miller, and Rivera (1984) took this a step further by restricting the various judges to

different performance domains. They did not see any " . . . overlap of the kinds of concerns, nature of evidence, and even the rigor of interpretation that is common to all groups" (p. 313). They believed, therefore, that questions posed for the different evaluators should be distinctly different.

Epstein (1985) proposed that the "the fairest evaluation will result when teachers are rated by multiple judges . . . who have clear interests in effective teaching" (p. 10). She suggested that the judgments cover multiple criteria that are recognized indicators of effective teaching.

Arguments have been made for the inclusion of parents in the process of teacher evaluation. Recent reform efforts call for increased parental involvement in education and increased empowerment of parents for decision making (Becher, 1984). Little empirical research is available, however, to indicate that parent participation in teacher evaluation is widespread.

Grandjean and Vaughn (1981) contended that "a parent and his or her child are mutually significant others" who influence each others' attitude formations (p. 275). In a study of high school effectiveness, they found that parents' perceptions of school were affected by their children's opinions, but not the reverse. They associated this outcome with the child regularly having more direct experience with the school than the parent.

Epstein (1985) maintained that parents may, in fact, be even more cognizant of teacher quality than principals. She undertook one of the most comprehensive efforts to investigate the potential contributions of parents in teacher evaluation by looking at some of the factors that influenced principals' and parents' ratings of teachers. Epstein determined that each emphasized different components of teaching. For example, principals were more aware of how well teachers performed on extra duties than about changes in classroom

practice, thus doing a good job on noninstructional tasks may have resulted in higher marks from principals than exemplary teaching. Parents, however, were more knowledgeable about special efforts the teacher made to help their children, and this had more impact on their ratings.

Peterson (1987) assessed teachers in one district to determine congruence among eight lines of evidence about job performance. He included competency tests, student achievement, professional involvement, and experience, as well as ratings by students, administrators, peers, and parents. Each line of evidence utilized different components of teacher behaviors (i.e., parents were surveyed about the information teachers provide, students about classroom learning, and principals made recommendations about overall qualifications). The information from parents and students had the highest correlation, and both had a low correlation with the principals' data.

Peterson's (1989) findings corroborated Epstein's work (1985) with regard to ratings by parents of elementary students but left questions about the use of parent ratings for secondary teachers. Peterson concluded that "less contact and communication results in more global or halo ratings for the teachers of older students" (p. 247).

Among the arguments presented for including students as evaluators is the conviction that students are the primary consumers of the teacher's services. As direct recipients of the teaching/learning process, students are the major clients of teachers and they are in the key position to provide information about teacher effectiveness (Follman, 1992). Most importantly, students are the only one of the teacher's clients who have direct knowledge about classroom practices on a regular basis. Student perceptions of quality, therefore, may be

more meaningful to the teacher than judgments by any other client group (Peterson & Kauchak, 1982). Evidence regarding teacher and principal support for the practice of involving students is mixed, but some see student involvement as beneficial from the student's perspective as well as from the teachers (Aleamoni, 1981; Follman, 1992).

Peterson, Gunne, Miller, & Rivera (1984) questioned the ability of students at any level to make objective judgments about some evaluation issues. McGreal (1994) and Harris (1986) suggested that student involvement in teacher assessment should be limited to descriptions of life in the classroom, rather than ratings of teacher worth. Peterson (1987) found that principals' ratings of teachers were higher than parents or students; however, the ratings in his study focused on different aspects the teachers' job. There is some agreement that the practice of involving students in teacher evaluation should be restricted to formative evaluation (Aleamoni, 1987), but the evidence is not conclusive.

Arguments have also been made for including opportunities for self-assessment among the strategies used to determine teacher success (Darling-Hammond, Wise, & Pease, 1983). Carroll (1981) stated that self-evaluation can provide useful data by offering "information and perspectives that may be unavailable from other sources" (p. 180) but it has generally been restricted to appraisals designed to foster teacher improvement, because objectivity is questionable, if promotion, tenure, or salary is tied to the evaluation. It has been suggested that the most effective use of self-ratings came when they were compared with other data sources.

Summary

The call for multiple and variable lines of evidence about the job performance of a teacher has many proponents (Scriven, 1994; Harris, 1986; Darling-Hammond, Wise, & Pease, 1983; Peterson, 1987; Stanley & Popham, 1988). It is illogical to expect administrators to be proficient in all of the roles required of them in schools today. It is also unrealistic to expect teachers to make desirable changes in practice unless they receive useful and helpful information about the quality of their work from authoritative and respected sources.

Empirical evidence is needed, therefore, to determine what multiple judges might contribute to the process of teacher evaluation (McGreal, 1994; Peterson 1987; Epstein, 1985). Such information can improve overall data quality and give teachers a basis for selecting information sources that will provide feedback for personal and professional improvement. This will result in better feedback for teachers (Tucker, Bray, & Howard, 1989) and will relieve the principal of the burden of being the sole judge of instructional quality (Peterson, 1984).

The use of multiple judges may also eliminate some of the negative effects that result from the evaluation process (Demming, 1986; Frase & Streshly, 1994). If teachers participate in the process, are permitted choice in the selection of data sources, and are comfortable with the procedures, outcomes may include better data as well as better use of the data.

PURPOSE OF THE STUDY

The call for multiple judges for the evaluation of teachers is based on the assumption that various client groups have different insights about an individual

teacher's performance. If this assumption is valid, then the different groups should provide different ratings of teachers on identical criteria of teaching effectiveness. However, there exists little empirical evidence that the use of different or multiple judges results in teacher evaluations that are different from those of building administrators. Several authors (Epstein, 1985; Peterson and Mitchell, 1985; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984) support Peterson's (1987) assertion, "No single line of evidence is sufficiently reliable, works for all practitioners, addresses all that a teacher does, or is agreed to by all educators" (p. 312). Most of the prior research, however, has centered on the use of data from various groups, with each group using a different set of criteria to assess teaching performance. None of these studies has looked at assessments by multiple judges using the same criteria of teaching effectiveness.

Until empirical evidence confirms that different groups provide different ratings on common criteria, the issue of employing multiple judges is of little importance in assessments of teaching. For example, if the different groups provide identical ratings, the use of more than one judge will provide only redundant information.

The purpose of this study was to compare teacher ratings from four judges on a common evaluation instrument. The data were used not only to determine whether or not the mean ratings of the four groups differed, but also, in the case of significant differences, to ascertain in which of six subcategories (i.e., Communication, Classroom Environment, Homework, Grading, Instruction, and Interpersonal Relationships) the differences occurred.

METHODOLOGY

Subjects

Samples representing four different populations were selected. These populations included: (1) teachers, (2) parents, (3) students, and (4) principals. The subjects in this study were the teachers in a large suburban school system and the principals, students, and parents associated with these teachers.

Teachers

A sample of 108 teachers was selected. The teachers in this sample were randomly chosen from a population of 1,610 teachers (37.5 percent of the teaching force) who volunteered to participate in a client satisfaction project conducted in the school district where they were employed. The teachers who comprised the project population agreed to solicit feedback about their job performance from the parents of their students. The teachers included representatives from all grade levels and all subjects.

The sample size (i.e.,108 teachers) was determined by specifying that fifteen teachers in each of seven categories should be included. Inclusion was limited to elementary grade level teachers and secondary core subject teachers (i.e., language arts, mathematics, science, and social studies).

Grade three was set as the lower limit for participation based on analysis of the reading level of the survey questions. It was determined that students in grade three and above could read and comprehend all items; however, interpretation was made available to younger students upon request. Since the school district was in the process of restructuring from a junior high school

configuration (i.e., K-6, 7-9, 10-12) to a middle school configuration (i.e., K-5, 6-8, 9-12), some grade six teachers in the study were assigned to an elementary school and some were assigned to a middle school.

Students

The students in this study were selected by choosing a random sample from each teacher's classroom. For elementary school teachers, the sample included 15 of the students assigned to each teacher. For middle school and high school core subject teachers, the sample included 15 of the students assigned to the teache.

Parents

The parents of students assigned to each teacher were selected. For elementary teachers, all of the parents of their assigned students were selected. For middle school and high school teachers, the parents of all students assigned to the teacher's second period class were chosen. A minimum of 20 parents were identified for each teacher, whenever possible. The parents were selected from the same class(es) from which the students were selected. Since all responses were anonymous, it was not possible to match the completed forms from parents with the completed forms of their own children.

Principals

The building principal of each teacher in the study was selected. Some principals rated more than one teacher.

Instrumentation

The instrument used in this study was a questionnaire that was designed by a group of teachers in a large suburban school district. These teachers were charged with the development of a customer (or client) satisfaction plan. The intent of this plan was to gather data that would let teachers know parent perceptions of their classroom effectiveness.

Participation in the project was voluntary and open to all teachers in the school system (n=4,285). All data were returned to the teachers confidentially and were used for teacher self-improvement.

Scoring

Teacher ratings were calculated as follows: responses under *Strongly Agree* were assigned a value of 4, those under *Agree* a value of 3, those under *Disagree* a value of 2, and those under *Strongly Disgree* a value of 1. Since the mean is an average of these values, it could range from a high of 4.0 (if all responses were *Strongly Agree*) to a low of 1.0 (if all responses were *Strongly Disagree*). Responses under *No Answer* and items left blank did not affect a teacher's rating. A grand mean or average of all of the item means was also calculated. It was this grand mean that provided the teacher ratings. No ratings were calculated for the subcategories.

Data Analysis

The questions posed in this study were addressed by testing the following null hypotheses:

Hypothesis 0_1: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on Classroom Environment.

Hypothesis 0_2: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on Homework.

Hypothesis 0_3: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on Grading.

Hypothesis 0_4: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on Communication.

Hypothesis 0_5: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on Instruction.

Hypothesis 0_6: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on Interpersonal Relationships.

Hypothesis 0_7: There is no significant difference among the mean ratings of teachers, principals, students, and teachers on the total score.

The data collected in this study were analyzed to provide statistical tests of the stated hypotheses. In each analysis, the F value for the differences between ratings was checked for statistical significance at the .05 level. Initially, the information was cast in the format displayed in Table 1.

Table 1

Format for Data Analysis for Each Participating Teacher

	Parents	Students	Principal	Teacher
Teacher 1	N=25	N=15	N=1	N=1
Teacher 2	N=25	N=15	N=1	N=1
•				
•				
•				
Teacher K	N=25	N=15	N=1	N=1

The data from the parents and students were converted to means. These means, the principal ratings and the teacher's self-assessments were analyzed by utilizing a one-way repeated measures analysis of variance design. This design is depicted in Table 2.

Table 2

Format for Data Analysis of Mean Ratings

	Parents	Students	Principal	Teacher
Teacher 1	N=1 (X for all)	N=1 (X for all)	N=1	N=1
Teacher 2	N=1 (X for all)	N=1 (X for all)	N=1	N=1
•				
•				
•				
Teacher K	N=1 (X for all)	N=1 (X for all)	N=1	N=1

In this study, the one-way analysis of variance design was utilized seven times, once for each subscore and once for the total of the six subscores. In each analysis, the F value for the differences between ratings was checked for statistical significance at the .05 level. If the null hypotheses were rejected in any of the seven analyses, the Tukey post-hoc procedure were used to determine the locations of the significant mean differences.

Additionally, the data from Table 2 were analyzed by computing Pearson correlations among the four ratings. Again, the correlations were computed for each subcategory as well as the total of all categories. The correlations were checked for significant differences from a population correlation of zero.

In this study, the one-way analysis of variance design was utilized seven times, once for each subscore and once for the total of the six subscores. In each analysis, the F value for the differences between ratings was checked for statistical significance at the .05 level. If the null hypotheses were rejected in any of the seven analyses, the Tukey post-hoc procedure were used to determine the locations of the significant mean differences.

Additionally, the data from Table 2 were analyzed by computing Pearson correlations among the four ratings. Again, the correlations were computed for each subcategory as well as the total of all categories. The correlations were checked for significant differences from a population correlation of zero.

ANALYSIS OF THE DATA

This section contains the results of the statistical analyses performed on the data collected in this investigation. Both descriptive statistics and the results from statistical tests are presented.

One hundred and eight teachers (42 elementary school teachers, 26 middle school teachers, and 40 high school teacher) were asked to participate in this study. Complete information was obtained for 93 teachers. The subjects included 33 elementary school teachers, 23 middle school teachers, and 37 high school teachers. These teachers represented the following subjects or grade levels:

(1) third grade teachers (n=11)

(2) fourth grade teachers (n=8)

(3) fifth grade teachers (n=10)

(4) sixth grade teachers (n=10)

(5) secondary science teachers (n=14)

(6) secondary language arts teachers (n=12)

(7) secondary social studies teachers (n=15)

(8) secondary mathematics teachers (n=13)

The data presented in this chapter are based on teacher self-assessments as well as ratings made by students, parents, and principals associated with these 93 teachers. The number of parents associated with each teacher ranged from six to eighteen and the number of student respondents per teacher ranged from thirteen to fifteen. Additionally, each teacher's principal completed the survey.

A subscore was determined for each teacher in the areas of Classroom Environment, Grading, Homework, Communication, Instruction, and Interpersonal Relationships. Also, a total mean score was calculated representing an overall average of the six subcategories. Each of the scores was expressed as the average item score per area. After student means and parent means were determined, average ratings and standard deviations were computed for each of the four groups of raters for each of the six subscores and total scores.

In general, parents and students rated elementary teachers higher than middle school or high school teachers. No significant differences were noted between the groups of teachers on principal ratings or teacher self-assessments.

The null hypotheses stated in this study were tested by a comparison of means based on the total sample of 93 teachers. Likewise, questions concerning relationships among the ratings of parents, principals, students, and teachers were addressed by calculating correlation coefficients on the total sample of teachers.

The results of the analyses on the subscore of Classroom Environment are presented in Tables 3 and 4.

Table 3

Means and Standard Deviations for Each Group on the Subcategory of Classroom Environment

Variable	Students	Parents	Teachers	Principals
$\bar{X}$	3.36	3.44	3.68	3.75
Range	2.3—4.0	2.3—4.0	3.0—4.0	3.0—4.0
SD	.34	.30	.40	.40

A one-way repeated measures analysis of variance conducted on these means produced an F (3, 276) of 32.22 ($p<.0001$). Tukey HSD post hoc comparisons indicated the population means for students and parents differed significantly from those of teachers and principals; thus, Hypothesis 0_1 was rejected.

Pearson correlation coefficients among the ratings of the four groups are given in Table 4.

Table 4

Correlations Among the Groups on Classroom Environment Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.510*	1.000		
Teachers	.228*	.045	1.000	
Principals	.341*	.046	.177	1.000

* $p < .05$ two-tailed probabilities

Significant correlations were noted between the ratings of students and the ratings of parents, teachers, and principals. None of the correlations among the ratings of parents, teachers, and principals was statistically significant for the Classroom Environment subscore.

Means and standard deviations for the area of Grading are presented in Table 5.

Table 5

Means and Standard Deviations for Each Group on the Subcategory of Grading

Variable	Students	Parents	Teachers	Principals
$\bar{X}$	3.22	3.36	3.61	3.64
Range	2.3-4.0	2.3-4.0	3.0-4.0	3.0-4.0
SD	.31	.30	.36	.41

A one-way repeated measures analysis of variance produced an $F(3, 276)$ of 37.32 ($p<.0001$). Tukey HSD post hoc comparisons indicated the population means for students differed significantly from those of all of the other data sources (i.e., students differed from parents, principals, and teachers) and the mean of parent ratings differed from the mean principal rating. Based on these results, hypothesis Hypothesis 0_2 was rejected.

Pearson correlation coefficients among the ratings of the four groups on the Grading subscore were calculated. The correlations are given in Table 6.

Table 6

Correlations Among the Groups on Grading Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.471*	1.000		
Teachers	.134	.037	1.000	
Principals	.274*	.056	.056	1.000

* p <.05 two-tailed probabilities

Significant correlations were found between ratings of students and those of parents and principals. None of the other correlations was statistically significant.

The analyses for the area of Homework are presented in Table 7 and Table 8.

Table 7

Means and Standard Deviations for Each Group on the Subcategory of Homework

Variable	Students	Parents	Teachers	Principals
$\overline{X}$	3.29	3.32	3.62	3.46
Range	2.5—3.8	2.4—4.0	2.8—4.0	2.6—4.0
SD	.30	.26	.39	.45

The one-way repeated measures analysis of variance resulted in an F (3, 276) of 19.23 ($p<.0001$). Significant differences were found between the populations means of students and those of teachers and principals and the population mean for parents differed significantly from the mean for teachers. Hence, Hypothesis 0_3 was rejected.

Pearson correlation coefficients among the ratings of the four groups were calculated. The correlations are given in Table 8.

Table 8

Correlations Among the Groups on Homework Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.528*	1.000		
Teachers	.207	.146	1.000	
Principals	.208*	.116	.027	1.000

* $p < .05$ two-tailed probabilities

As in the previous analyses, the only significant correlations involved student ratings. These ratings correlated significantly with the rating of parents, teachers, and principals.

The results of the analyses for the area of Communication are presented in Tables 9 and 10.

Table 9

Means and Standard Deviations for Each Group on the Subcategory of Communication

Variable	Students	Parents	Teachers	Principals
$\bar{X}$	3.23	3.30	3.56	3.56
Range	2.4—3.8	2.3—3.9	2.9—4.0	2.3—4.0
SD	.34	.33	.36	.44

Results of a one-way repeated measures analysis of variance conducted on these means produced an F (3, 276) of 23.29 ($p<.0001$). Tukey HSD post hoc comparisons indicated the population means for students and parents differed significantly from those of teachers and principals. Based on these results, Hypothesis 0_4 was rejected.

Pearson correlation coefficients among the ratings of the four groups were calculated. The correlations are given in Table 10.

Table 10

Correlations Among the Groups on Communication Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.473*	1.000		
Teachers	.037	.112	1.000	
Principals	.262*	.010	-.024	1.000

* $p < .05$ two-tailed probabilities

Again, significant correlations were present between student ratings and the rating of parents and principals.

The analyses of the area of Instruction are presented in Table 11 and Table 12.

Table 11

Means and Standard Deviations for Each Group on the Subcategory of Instruction

Variable	Students	Parents	Teachers	Principals
$\overline{X}$	3.26	3.39	3.57	3.65
Range	2.4—3.8	2.6—3.8	2.9—4.0	2.7—4.0
SD	.35	.25	.30	.37

A one-way repeated measures analysis of variance conducted on these means produced an F (3, 276) of 35.11 ($p<.0001$). Tukey HSD post hoc comparisons indicated the population means for students differed significantly from those of all of the other data sources (i.e., students differed from parents, principals, and teachers) and the population means for parents differed significantly from those of principals and teachers. Hypothesis 0_5 was rejected.

Pearson correlation coefficients among the ratings of the four groups were calculated. The correlations are given in Table 12.

Table 12

Correlations Among the Groups on Instruction Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.480*	1.000		
Teachers	.150	.133	1.000	
Principals	.424*	.015	.030	1.000

* p <.05 two-tailed probabilities

The correlations between student ratings and those of parents and principals were statistically significant.

The results for the area of Interpersonal Relationships are presented in Table 13.

Table 13

Means and Standard Deviations for Each Group on the Subcategory of Interpersonal Relationships

Variable	Students	Parents	Teachers	Principals
$\bar{X}$	3.19	3.38	3.59	3.65
Range	2.0—3.9	2.6—3.8	2.8—4.0	1.8—4.0
SD	.42	.28	.32	.42

A one-way repeated measures analysis of variance resulted in an F (3, 276) of 39.96 (p<.0001). Tukey HSD post hoc comparisons indicated the

population means for students differed significantly from those of all of the other data sources (i.e., students differed from parents, principals, and teachers) and the population means of parents differed from the means of principals and teachers. These findings indicated that Hypothesis 0_6 was rejected.

Pearson correlation coefficients among the ratings of the four groups were calculated. The correlations are given in Table 14.

Table 14

Correlations Among the Groups on Interpersonal Relationships Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.540*	1.000		
Teachers	.110	.094	1.000	
Principals	.347*	-.030	.134	1.000

* $p < .05$ two-tailed probabilities

As was found for the other subscores, ratings from students correlated significantly with the ratings of parents and principals.

Finally, the total score or mean of all the subcategories was analyzed. These results are summarized in Tables 15 and 16.

Table 15

Means and Standard Deviations for Each Group on the Total Score

Variable	Students	Parents	Teachers	Principals
$\bar{X}$	3.25	3.36	3.58	3.60
Range	2.4—3.7	2.7—3.8	3.0—4.0	2.6—4.0
SD	.33	.25	.27	.36

An F (3, 276) of 41.99 ($p<.0001$) was obtained from the one-way repeated measures analysis of variance conducted on these means. Tukey HSD post hoc comparisons indicated the population means for students and parents differed significantly from those of teachers and principals; therefore, Hypothesis 07 was rejected.

Pearson correlation coefficients among the ratings of the four groups were calculated. The correlations are given in Table 16.

Table 16

Correlations Among the Total Scores

Rater	Students	Parents	Teachers	Principals
Students	1.000			
Parents	.544*	1.000		
Teachers	.139	.098	1.000	
Principals	.345*	.012	.079	1.000

* $p < .05$ two-tailed probabilities

The only significant correlations were those between the ratings of students with the ratings of both parents and principals.

Correlation coefficients for elementary school, middle school, and high school teachers can be found in Appendicies K-Q (pages 123-163).

A graph depicting the means of the four data sources in each subcategory is presented in Figure 1 (page 27).

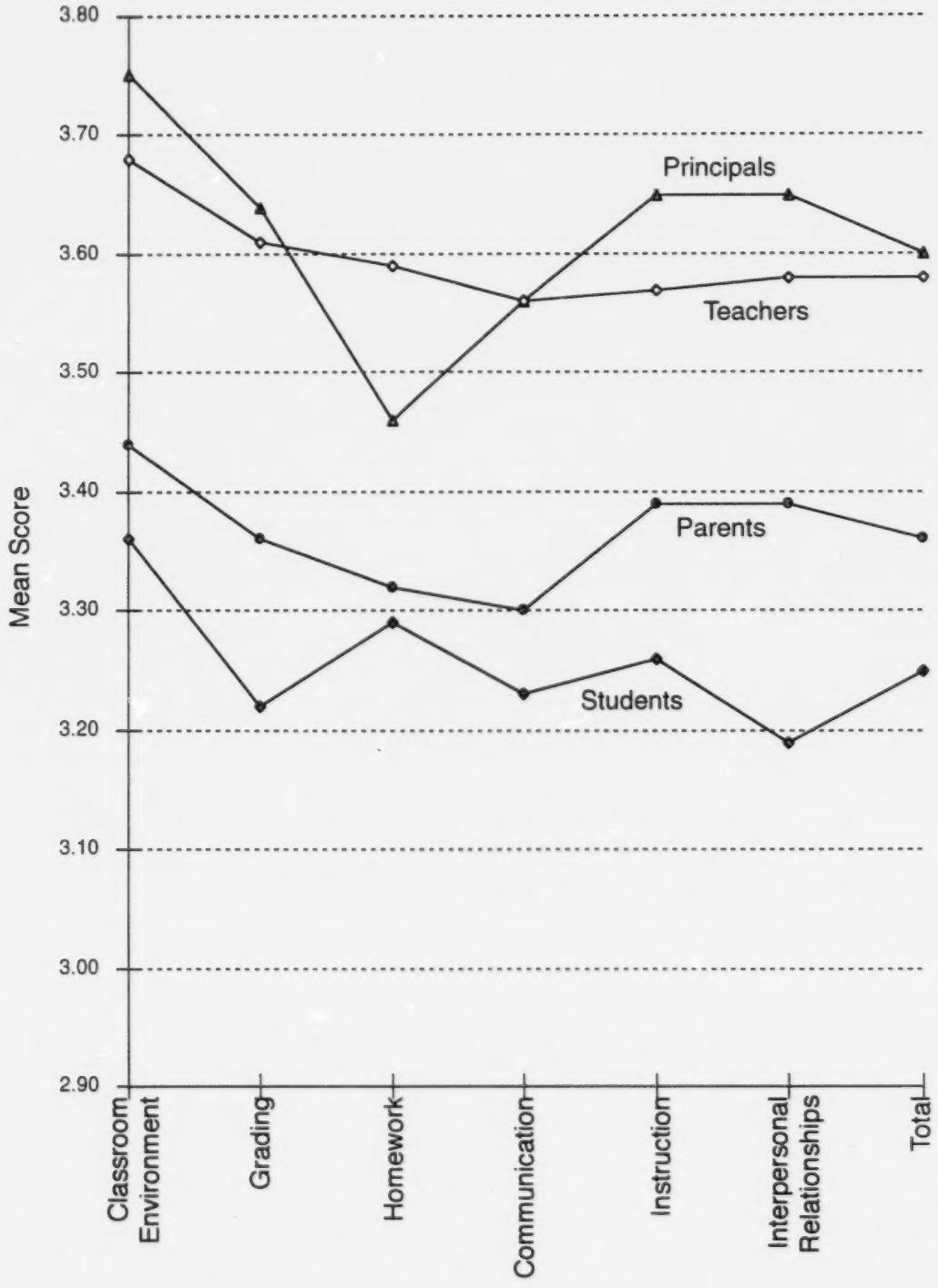

Figure 1. Graph Comparing Mean Scores of Four Data Sources

DISCUSSION AND CONCLUSIONS

Discussion

In general, the ratings assigned the 93 teachers in this study were very high. All subscore item averages were greater than 3.0. This indicates that for each area the average response was between "Agree" and "Strongly Agree" and was, therefore, a positive statement concerning a teacher's performance.

The highest ratings were given by principals. Principals' average ratings for teachers ranged from 3.46 for Homework to 3.75 for Classroom Environment. Only on the area of Homework did teachers receive a higher average rating than that given by principals. In this one area, teachers rated themselves slightly higher than did principals (3.59 versus 3.46).

The second highest ratings were given by teachers. However, the analysis of variance and the post hoc procedures indicated that only on the Homework subcategory did the population means provided by teachers differ significantly from those assigned by principals. The absence of significant differences in the other areas may be due to the fact that both sets of ratings approached the upper limit of four. The fact that both principals and teachers give teachers high ratings has been noted in the literature (Haefele, 1992; Frase & Streshley, 1994; Centra, 1972). But, the literature suggested that teachers' self-ratings are generally higher than those given by administrators (Payne & Hulme, 1988).

Since principals were once classroom teachers, they are likely to be more empathetic towards the job of a teacher than parents or students and this may account for their high assessment of the teachers in this study. Bridges

(1986) contended that principals are also aware that they do not devote enough time to teacher evaluation to do an adequate job and consequently are reluctant to be critical of teacher performance. Another plausible explanation for the higher ratings from teachers and principals is that these two groups possess more knowledge about the job requirements and working conditions for the teacher than other raters and it is this information that generates such high assessments. Epstein (1985) speculated "principals may give greater consideration than parents to demographic demands on teachers that require good teaching and classroom management skills" (p. 5).

The fact that principal ratings on the Homework variable were lower than those of teachers may have occurred because principals are often the mediators when parents encounter problems related to their children's home assignments. The principals could, therefore, be more aware of negative client perceptions in this area than in some of the other subcategories.

Even though both principals and teachers gave teachers high ratings, the correspondence between the two sets of ratings was quite low. None of the correlation coefficients between the two groups differed significantly from zero. This lack of correspondence may have been influenced by the relatively small variances associated with the rankings, but, since other sets of ratings with equally small standard deviations were significantly correlated, additional factors must be considered.

The literature suggests that principals' ratings may be influenced by different factors than those governing teacher self-assessments. For example, Scriven (1981) contended that it is difficult for the principal to separate the teacher's value to the institution from responsibilities that are linked to

classroom performance and Wood (1992) claimed that assessments are often based on preexisting conceptions. Others (Root & Overly, 1990; Haefele, 1992) have noted that principals have little direct information on teachers' classroom performance. Thus, the findings may indicate that, in general, principals' and teachers' opinions of teacher competence may differ. This notion is supported in a study by Centra (1972) involving college faculty. Centra found a low correlation between self-ratings and ratings given by administrators.

The teacher ratings given by parents and students, while quite high, were significantly lower than those given by teachers and principals on each of the six areas. This finding is, to a degree, supported by prior research. In a study involving high school teachers, Peterson and Yaakobi (1980) found that teacher self-assessments were higher than student ratings. Also, Centra (1972) reported that college faculty rated themselves higher than did their students. Conversely, Peterson (1987) reported average principal ratings of 3.92 and average parent ratings of 4.52. It should be noted, however, that the parents and principals in Peterson's study employed different criteria for rating teachers. This may explain the conflicting results in the present study.

Epstein (1985) stated that parents may be more cognizant of teaching effectiveness than principals. She also suggested that principals may consider other duties that the teacher performs or assess the teacher's leadership skills when making their ratings, but the parents are more apt to consider the teachers' interactions with students. If true, this could explain the differences between the parents' and the principals' ratings.

The results obtained in this study support the widely held view and the findings of a prior empirical study that parents' perceptions of schools are

based on the opinions of their children (Clark, 1987). In an investigation of parent and student opinions towards school and school programs, Grandjean and Vaughn (1981) concluded that parents' perceptions of school were affected by their children's opinion, but not the reverse. Though causality cannot be inferred, the correlations between parent and student ratings were higher on every area than the correlations between other groups employed in this study. These values ranged from .47 to .54, although the variances for the mean ratings were quite small. These correlations clearly indicate a relationship between student and parent perceptions of teacher performance. It is plausible that this connection is causal to the extent that parents form opinions toward teachers on the basis of information supplied by their children. This finding is supported by Peterson (1987). In his study, information from parents and students were significantly correlated and both had near zero correlations with principal ratings.

On three of the six areas, average student ratings were significantly lower than mean parent ratings. These areas were Grading, Instruction, and Interpersonal Relationships. These differences and the fact that student ratings were slightly lower on the other three scales indicated that students are more negative toward teacher performance than parents.

The biggest difference between student and parent ratings occurred on the subscore of Interpersonal Relationships (e.g., courtesy, respect). It may be that since teachers have far fewer interactions with parents than they do with students, they exhibit only the most positive aspects of their personalities during the occasional parent contacts. Since no information about teacher-parent interactions was collected, it is not possible to determine if the parents

in this study had ever had face-to-face meetings or even telephone conversations with the teachers they rated.

Whether the differences between parents and students are related to the validity of either set of ratings cannot be determined by the data generated in the present study. Several writers (Larson, 1984; Peterson & Kauchak, 1982; Follman, 1992) have contended that students possess more direct knowledge of teacher behavior than administrators or parents. Lower student ratings may be the result of this knowledge.

In addition to the significant correlations found between mean student and mean parent ratings, average student ratings also consistently correlated with the principals' ratings of teachers. These values ranged from .20 to .42 across the six areas. In only two of the six areas were significant correlations noted between average student ratings and teacher ratings. These two values were .28 (Environment) and .21 (Homework). Why the significant correlations found in these data involve average student ratings is not readily apparent. However, these findings are consistent with the suppositions that: (1) teacher self-perceptions are subjective and consequently may not be consistent with their actual performance (Peterson & Kauchak, 1982); (2) students possess more direct knowledge of teacher performance than other groups (Follman, 1992); (3) when this knowledge is shared with parents, parents opinions are altered (Grandjean & Vaughn, 1981); and (4) principals also have some knowledge of classroom behavior and thus principal ratings correlate to some degree with student ratings.

Conclusions

This study addressed the questions of differences between student, parent, principal, and teacher ratings of teaching performance. Prior research in this area has generally involved different groups (most often students and principals) who rated different dimensions of teaching behaviors. In this investigation, all four groups rated teacher performance on the same instrument. The results of the statistical analyses support the following conclusions:

1. Teachers are viewed in a positive light by all four groups.
2. Students and parents have more negative perceptions about teacher performance than do principals or teachers.
3. While teachers and principals rate teachers highly, there is little agreement between their rank ordering of teachers.
4. There exists a significant degree of agreement between students' and parents' rank ordering of teachers on teaching performance.
5. Teachers are viewed less positively by students than by parents.

On a more general level, the five conclusions indicate that the use of multiple judges may provide unique perspectives on teacher performance. This was consistent with Epstein's (1985) assertion, "Because there is not a single set of skills that perfectly define effective teaching, measures of many aspects of teaching by multiple judges are likely to yield the fairest and most comprehensive evaluation of teachers" (p. 8).

Reference List

Aleamoni, L. (1981). Student ratings of instruction. In J. Millman (Ed.). Handbook of teacher evaluation, (pp. 110-145). Beverly Hills: Sage.

Becher, R. M. (1984). Parent involvement: A review of research and principles of successful practice. (Report No. PS 014 563). Urbana, IL: National Institute of Education. (ERIC Document Reproduction Service No. ED 247 032)

Bridges, E. (1986). The incompetent teacher. Philadelphia, PA: Falmer Press.

Carroll, J. G. (1981). Faculty self-evaluation. In J. Millman (Ed.). Handbook of teacher evaluation, (pp. 180-200). Beverly Hills: Sage.

Centra, J. (1972). Two studies on the utility of student ratings for instructional improvement. Princeton, NJ: Educational Testing Service.

Collins, A. (1991). Portfolios for biology teacher assessment. Journal of Personnel Evaluation in Education, 5, 147-267.

Darling-Hammond, L., Wise, A. E., & Pease, S .R. (1983). Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research, 53(3), 285-328.

Deming, W. E. (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology.

Dixon, R. G. D. (1994). Future schools: And how to get there from here. Phi Delta Kappan, 75, 360-365.

Ellett, C., & Garland, J. (1987). Teacher evaluation practices in our largest school districts: Are they measuring up to "state-of-the-art" systems. Journal of Personnel Evaluation in Education, 1, 69-92.

Epstein, J. (1985). A question of merit: Principals' and parents' evaluation of teachers. Educational Researcher, 14(7), 3-10.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. The High School Journal, 75, 168-178.

Frase, L., & Streshly, W. (1994). Lack of accuracy, feedback, and commitment in teacher evaluation. Journal of Personnel Evaluation in Education, 1, 47-57.

Grandjean, B., & Vaughn, III, E. (1981). Client perceptions of school effectiveness: A reciprocal causation model for students and their parents. Sociology of Education, 54, 275-290.

Haefele, D. (1981). Teacher interviews. In J. Millman (Ed.). Handbook of teacher evaluation, (pp. 41-57). Beverly Hills: Sage.

Haefele, D. (1992). Evaluating teachers: An alternative model. Journal of Evaluation in Education, 5, 335-345.

Haefele, D. (1993). Evaluating teachers: A call for change. Journal of Personnel Evaluation in Education, 7, 21-31.

Harris, B. (1986). Developmental teacher evaluation. Boston: Allyn and Bacon.

Larson, R. (1984). Teacher performance evaluation—What are the key elements? NASSP Bulletin, 68(469), 13-18.

Manning, R. C. (1988). The teacher evaluation handbook. Englewood Cliffs, NJ: Prentice Hall.

McGreal, T. L. (Speaker). (1994). The next generation of teacher evaluation. (Cassette Recording No. 94-4625). Chicago: Association for Supervision and Curriculum Development.

Medley, D. M., Coker, H., & Soar, R. S. (1984). Measurement -based evaluation of teacher performance: An empirical approach. New York: Longman.

Payne, D. A., & Hulme, G. (1988). The development, pilot implementation, and formative evaluation of a grass-roots teacher evaluation system: Or, the search for a better lawnmower. Journal of Personnel Evaluation in Education, 1, 365-372.

Peterson, K. (1984). Methodological problems in teacher evaluation. Journal of Research and Development in Education, 17(14), 62-70.

Peterson, K. (1987). Teacher evaluation with multiple and variable lines of evidence. American Educational Research Journal, 24, 311-317.

Peterson, K. (1989). Parent surveys for school teacher evaluation. Journal of Personnel Evaluation in Education, 2, 239-249.

Peterson, K., Gunne, G., Miller, P., & Rivera, O. (1984). Multiple audience rating form strategies for student evaluation of college teaching. Research in Higher Education, 20(3), 309-321.

Peterson, K., & Kauchak, D. (1982). Teacher evaluation: Perspectives, practices, and promises. (Report No. SP 022 900). Salt Lake City, UT: Center for Educational Practice. (ERIC Document Reproduction Service No. ED 233 996)

Peterson, K., & Mitchell, A. (1985). Teacher-controlled evaluation in a career ladder program. Educational Leadership, 43(3), 44-47.

Peterson, K., & Yaakobi, D. (1980). Israeli science students and teacher perceptions of classroom role performance: Concepts, reports, and adequacy. Science Education, 64, 661-669.

Root, D., & Overly, D. (1990). Successful teacher evaluation—key elements for success. NASSP Bulletin, 74(527), 34-38.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.). Handbook of teacher evaluation, (pp. 244-271). Beverly Hills: Sage.

Scriven, M. (1994). Duties of the teacher. Journal of Personnel Evaluation in Education, 8, 151-184.

Stanley, S., & Popham, W. (Eds.). (1988). Teacher evaluation: Six prescriptions for success. Alexandria, VA: Association for Supervision and Curriculum Development.

Stodolsky, S. (1984). Teacher evaluation: The limits of looking. Educational Researcher, 13(9), 11-18.

Tucker, N. A., Bray, S. W., & Howard, K. C. (1989). Using a client-centered approach in the principal's evaluation of counselors. Journal of Personnel Evaluation in Education, 2, 335-353.

Wise, A. E., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1984). Teacher evaluation, A study of effective practices. Santa Monica, CA: Rand Corporation.